JOURNAL OF SMALL BUSINESS MANAGEMENT
https://doi.org/10.1080/00472778.2022.2082453

Predicting New Venture Gestation Outcomes With Machine Learning Methods

Paris Koumbarakis[a] and Thierry Volery[b]

[a]University of St Gallen, Swiss Institute for Small Business & Entrepreneurship, Switzerland; [b]Zurich University of Applied Sciences, School of Management & Law, Switzerland

ABSTRACT

This study explores the use of machine learning methods to forecast the likelihood of firm birth and firm abandonment during the first five years of a new business gestation. The predictability of traditional logistic regression is compared with several machine learning methods, including logistic regression, k-nearest neighbors, random forest, extreme gradient boosting, support vector machines, and artificial neural networks. While extreme gradient boosting shows the best overall model performance, neural networks provide good results by correctly classifying entrepreneurs who have not abandoned their business venture in the early stage of the gestation process. In addition, this study provides valuable insights in relation to the start-up activities leading to firm emergence. Entrepreneurs who perform a greater number of activities and who can orchestrate them at the right rate, concentration, and time are more likely to successfully launch a new business venture.

KEYWORDS

New venture creation; forecasting; machine learning

Introduction

Launching a new business venture requires a lot of time and effort. According to one estimate, US business angels invested more than $26 billion into start-ups in 2018 (WBAF, 2020) and the time entrepreneurs devote to starting new firms amounts to 2.7% of total paid work (Reynolds & Curtin, 2011). The related sunk costs in start-ups are correspondingly high, considering that the majority of entrepreneurs abandon their business idea in the first five years of a new business creation (Reynolds, 2017). Against this backdrop, investors and service providers continuously face the decision to put extra time, effort, and money to support entrepreneurs and their fledgling business ventures; but how do they know whether these entrepreneurs will launch a profitable business venture, quickly abandon their business idea, or spend years pottering about a business idea that never gets traction? Accurate prediction models of venture gestation could help investors and other stakeholders optimize their resource allocation.

CONTACT Thierry Volery thierry.volery@zhaw.ch Zurich University of Applied Sciences, School of Management & Law, Winterthur, Switzerland

In this study, we contribute to the literature on new venture creation by tackling two research questions: (1) What model(s) best predict firm birth and firm abandonment? And (2) What are the single conditions leading to firm emergence? To this end, we use artificial intelligence (AI) techniques to forecast the likelihood of firm emergence and firm abandonment during the first five years of a new business gestation. By exploring the combinations of single conditions, and by using machine learning methods to forecast firm emergence, we aim to generate new insights on the start-up process and not just test theory. This approach is warranted because despite the multitude of research on firm emergence, the extant literature on start-up processes remains fragmented (Davidsson & Gruenhagen, 2020). In addition, previous research seems to have reached an empirical dead-end in trying to identify what combinations of single conditions are more likely to explain firm emergence (Arenius et al., 2017). The need to uncover the necessary conditions for firm emergence is thus more actual than ever.

There exist many prediction models which have identified start-up success and failure factors across different countries (for example, Lussier & Claudia, 2010; Mayr et al., 2020). However, these models have several limitations. First, they relate to the later stage of the start-up process, rather than the whole gestation period (Reynolds, 2017). For example, past studies often adopted a firm perspective, drawing on data collected from owners of already established or liquidated companies (Lussier, 1995). Second, most prediction models have traditionally drawn on linear or logistic regression methods. Recently, entrepreneurship scholars have begun using machine learning (AI) techniques to address multiple research questions related to new venture dynamics (Antretter et al., 2019; Van Witteloostuijn & Kolkman, 2019; Weinblat, 2018). These AI techniques often outperform traditional regression-based models, providing higher prediction accuracy (Loureiro et al., 2018) and detecting ambiguity of interaction and nonlinear effects in input data (Gerasimovic & Bugaric, 2018). Third, AI models, specifically the supervised learning model types, provide great added value in predictive tasks since they are specifically designed for such purposes (Obschonka & Audretsch, 2019).

We provide a longitudinal, multifaceted perspective on the start-up process by drawing on a large set of harmonized international panel data used to consider multiple factors affecting firm emergence, including start-up activities and their evolution over time, as well as business, industry, and entrepreneurs' characteristics. In this context, AI can help reveal unexpected patterns in a data set and potential connections between otherwise unrelated issues, which in turn could serve as a basis for developing new theories in entrepreneurship (Lévesque et al., 2020; Obschonka & Audretsch, 2019). By

exploring the combinations of single factors using new AI methods to forecast firm emergence, we do not aim to test theories, but to garner important new insights into the field of entrepreneurship.

The remainder of this paper is structured as follows: the second section provides a review of the factors influencing the venture creation process and a rationale for the quantitative exploratory approach. The third section describes the methodology, including the data set, variables, and empirical methods. The fourth section details the results, followed by a discussion section and the conclusion.

Background

Factors influencing venture gestation outcome

The prediction of venture gestation outcome has been a central question in entrepreneurship research over the past two decades (Tornikoski & Newbert, 2007; Van Gelderen et al., 2005). Given the prevalence of the entrepreneur and the lack of financial information in the early stages of the business gestation process, past research has focused predominantly on nonfinancial elements such as the occurrence and sequence of gestation activities (for example, prototype development, market validation, team recruitment, raising capital) which may impact the likelihood of firm birth and firm abandonment (Burnaev et al., 2015; Newbert, 2005).

Empirical research in this field has mainly drawn from Panel Study of Entrepreneurial Dynamics (PSED) research programs or studies following a similar design (Davidsson & Gordon, 2012; Reynolds & Curtin, 2011). PSED provides reliable and generalizable data on the process of business formation. It includes information on the characteristics of the adult population attempting to start new businesses, the kinds of activities nascent entrepreneurs undertake during the business start-up process, and the proportion and characteristics of the start-up efforts that are launched. Even though the results regarding the factors that impact the gestation outcome are diverse, the following patterns have emerged from this stream of research (Table 1).

First, the type as well as the amount of gestation activities appear to impact the venture gestation outcome (Chwolka & Raith, 2012; Honig & Samuelsson, 2012). In fact, "what nascent entrepreneurs do may be more important than whom they are and what product-markets they intend to serve" (Tornikoski & Newbert, 2007, p. 313). Taking a closer look at single start-up activities, business planning and raising funds have been closely related to the likelihood of firm birth. Longitudinal studies suggest that business planning plays an important role in the gestation outcome (Liao & Gartner, 2007; Newbert & Tornikoski, 2012). Specifically, business planning facilitates goal attainment as it helps founders to undertake more valuable actions to develop their fledgling

Table 1. Predictors of firm emergence.

Category	Description	References
Gestation activities		
Business planning	Business plan prepared; financial projections prepared; effort made to define the market opportunity	Chwolka and Raith (2012); Delmar and Shane (2003); Liao and Gartner (2007); Newbert and Tornikoski (2012)
Seeking funding	Outside funding from institutions or people received; credit from suppliers established, own money invested into the business	Van Gelderen et al. (2005); Hechavarria et al. (2012). Liao et al. (2009)
Complexity measures: timing, rate, concentration	Timing: time at which activities are conducted; rate: number of activities undertaken over a period of time; concentration: temporal concurrency of different activities	Lichtenstein et al. (2007); Hopp and Sonderegger (2015)
Human and social capital		
Psychological characteristics	Goal commitment; self-efficacy,	Hechavarria et al. (2012); Khan et al. (2014),
Personal background	Industry experience; entrepreneurial experience; educational attainment; gender	Dimov (2010); Rotefoss and Kolvereid (2005); Tu et al. (2019)
Team characteristics	Team size; team experience; team composition	Chandler et al. (2005), Delmar and Shane (2006); Muñoz-Bullon et al. (2015) Steffens et al. (2012); Thiess et al. (2016)
Social capital	Entrepreneur's personal network; social network; strong ties vs weak ties in the process	Clough et al. (2019); Davidsson and Honig (2003); Hallen, 2008; Hallen et al. (2020)
Context		
Industry	Service vs manufacturing; technology-based vs non-technology-based ventures	Van Gelderen et al. (2005); Steffens et al. (2012); Liao and Welsch (2008)
Market dynamism	Level of competition in the industry; low-velocity vs moderate-velocity markets	Newbert (2005); Hopp and Sonderegger (2015)
Economic cycle	Boom vs bust; economic crisis	Giotopoulos et al. (2017); Vegetti and Adăscăliței (2017)

enterprises (Delmar & Shane, 2003). When it comes to start-up capital and funding, the percentage of ownership as well as external equity influence birth outcome (Hechavarria et al., 2012; Van Gelderen et al., 2005). For example, financial capital significantly decreases the odds of discontinuance (Liao et al., 2009). Furthermore, networks and activities that connect the nascent entrepreneur with others appears to positively impact firm birth (Newbert & Tornikoski 2012; Newbert et al., 2013). However, the impact of specific activities is inconsistent (Chwolka & Raith, 2012; Delmar & Shane, 2003) and even though some level of activity is needed, no single gestation activity appears necessary to achieve firm birth (Arenius et al., 2017; Shim & Davidsson, 2018).

Next to specific activity types, three factors time, rate, and concentration have been studied to assess the "complexity dynamics" (Lichtenstein et al. 2007) of the gestation process. There is some evidence that the timing of start-up activities (whether the bulk of the organizing activities is accomplished earlier or later during the start-up process), the rate (the number of start-up activities undertaken over a period of time) and the concentration (how closely

start-up activities are undertaken in relation to each other) has an impact on firm emergence (Hopp & Sonderegger, 2015; Lichtenstein et al. 2007). New ventures are more likely to emerge when entrepreneurs conduct gestation activities at a faster rate, in lower concentration, and with an average timing at a later stage in the gestation process.

Next to start-up activities, it is widely recognized that personal characteristics, and in particular personal agency such as self-directedness and self-efficacy, have a positive impact on firm emergence and a negative relationship with firm abandonment (Dimov, 2010; Hechavarria et al., 2012; Khan et al., 2014). In addition, personal background characteristics including entrepreneurial experience (Rotefoss & Kolvereid, 2005; Van Gelderen et al., 2005), industry experience (Dimov, 2010), education attainment (Hopp & Sonderegger, 2015), and age (Liao et al., 2009) can affect the start-up activities conducted and accelerate the speed toward business launch.

With regard to the founding team, initial team size (Chandler et al., 2005), team resource heterogeneity (Muñoz-Bullon et al., 2015) as well as balanced team experience (Delmar & Shane, 2006; Thiess et al., 2016), positively influence the firm birth. In the context of homogenous teams, also with regard to homogenous start-up experience, there exists a negative relationship toward firm performance in the long-term (Steffens et al., 2012). Although contradicting results subsist (Tornikoski, 2008), the influence of the team on the venture outcome has largely been validated.

Next to the human capital perspective, there is wide agreement that social capital (the resources embedded in entrepreneurs' personal networks) is critical for the emergence of new firms. For instance, network connections enable entrepreneurs to identify new business opportunities, marshal resources, and secure legitimacy from external stakeholders (Clough et al., 2019; Hallen, 2008). In their study on nascent entrepreneurship comparing individuals engaged in start-up activities with a control group of nonentrepreneurs in Sweden, Davidsson and Honig (2003) found that social capital variables were very strong and consistent predictors of firm emergence. Both bonding and social capital based on strong ties, such as having parents who owned businesses or close friends who owned businesses, and bridging social capital based on weak ties were found to be a good predictor of nascent entrepreneurship. Recent research on accelerator and entrepreneurship education programs (Hallen et al., 2020) has documented how nascent entrepreneurs interact and learn within an accelerator, further expanding their social network in the process. These programs are important mechanisms for nascent entrepreneurs to attract resources and to convey quality and legitimacy.

Context variables matter too. For example, the industry in which the venture evolves has an impact on the gestation outcome. Ventures within the service industry are likely to be more rapidly operational and profitable (Steffens et al., 2012; Van Gelderen et al., 2005). Market dynamism is another

important contextual variable. For example, the number of gestation activities for nascent entrepreneurs operating in low-velocity markets is greater than for nascent entrepreneurs operating in moderate-velocity markets (Newbert, 2005). The use of technology is a key determinant of market dynamism. Technology-based entrepreneurs typically create business ventures operating in more dynamic and uncertain environments, and they engage in more planning, legitimacy establishment and resource acquisition activities (Liao & Welsch, 2008). Economic crises tend to have a negative impact on the emergence of new business ventures through a drastic drop in demand for goods and services (Giotopoulos et al., 2017; Vegetti & Adăscăliței, 2017).

Rationale for an exploratory quantitative approach

Despite the plethora of research, the rich literature on firm emergence is surprisingly limited in volume and results remain fragmented (Davidsson & Gruenhagen, 2020). Scholars have recently suggested that no particular gestation activity is necessary to achieve firm birth and that only a low number of activities is necessary for reaching initial profits after 24 months of gestation (Arenius et al., 2017). Two main reasons explain this fragmentation. First, past studies have typically adopted one single perspective or theoretical anchor, such as creative agency (Hechavarria et al., 2012; Khan et al., 2014), planning theory (Honig & Samuelsson, 2012; Liao & Gartner, 2007; Newbert et al., 2013), and human capital (Hopp & Sonderegger, 2015; Muñoz-Bullon et al., 2015; Steffens et al., 2012), therefore shedding light on one aspect of the new venture creation process at a time. Second, as a corollary, previous studies of gestation activities have primarily been content with a partial understanding of organizing activities as sufficient conditions (Arenius et al., 2017). However, the relative importance of each activity and an understanding of which activities constitute necessary conditions for firm emergence has yet to be established. Third, because of the temporal heterogeneity of venture creation processes, scholars have often focused on the achievement of partial milestones, such as receiving external funding or generating first sales. As a result, "it has been very difficult to find either general patterns in or explanations for the entire sequence of gestation activities" (Davidsson & Gordon, 2012, p. 858).

Recognizing that prior research may have reached an empirical dead-end in trying to identify gestation activities as sufficient conditions for firm emergence (Arenius et al., 2017), we seek to explore the combinations of single conditions which are more likely to explain firm emergence. Our approach thus departs from most recent research on new venture gestation characterized by a proliferation of quantitative work aiming to extend or add new theoretical understanding, a procedure which could explain the slow rate of cumulative research progress in the field (Wennberg and Anderson 2020).

To this end, we use machine learning methods to forecast the likelihood of firm emergence and firm abandonment during the first five years of a new business gestation. By exploring the combinations of large sets of variables and adopting new methods to forecast firm emergence, we aim to generate new insights, rather than just testing theories. Our context of applying machine learning methods to a large data set (PSED) implies that we are engaging in exploratory data-driven empirical research (Coad & Srhoj, 2019; Wennberg and Anderson 2020) as a fact-finding exercise that could help trigger novel theories which run counter to existing ones or broaden the scope of existing ones by identifying variables and relationships from areas ignored in the past (Lévesque et al., 2020).

Methodology

Data set

This study draws on five longitudinal data sets from Panel Study of Entrepreneurial Dynamics (PSED) research program in four different countries: the US PSED I (1998–2004) and US PSED II (2005–2008), the Swedish Panel Study of Entrepreneurial Dynamics, the Comprehensive Australian Study of Entrepreneurial Emergence, and the Chinese PSED. These data sets have been harmonized into one data set which comprises 3537 nascent entrepreneurs (Reynolds et al., 2016).

PSED provides valid and reliable data on the process of business formation based on nationally representative samples of nascent entrepreneurs. Its design is based on a population screening interview to identify nascent entrepreneurs and a series of subsequent interviews to track their progress toward their business launch. In this study, entrepreneurs were tracked over a period of 60 months following their first identification as "nascent" in the screening interview. To be classified as nascent, entrepreneurs had to perform at least two gestation activities (for example, develop, a prototype, draft a business plan, register a business, open a bank account, recruit first employee, and so on) in the 12 months prior to the screening interview. Based on this entry point date, nascent entrepreneurs provided information about the completion of subsequent gestation activities and the outcome of the gestation process (that is, firm birth, firm abandonment, and ongoing gestation) in three-month intervals. As a result of this selection procedure, our data set consisted of 1457 nascent entrepreneurs.

In terms of gestation outcome, PSED defines firm birth as the presence of monthly profits that cover expenses and owner salaries, firm abandonment as having stopped working on the business idea, and ongoing gestation as entrepreneurs still pursuing their business idea, having neither set up a profitable new firm nor abandoned their business idea (Reynolds et al.,

2016). In this study, we focus on two gestation outcomes: firm birth and firm abandonment. It should be noted that firm abandonment occurs during the gestation process before the firm emerges and becomes profitable. We discarded nascent entrepreneurs in ongoing gestation. These individuals have often been characterized as "dilettante dreamers" or "hobbyists" (Davidsson & Gordon, 2012; Reynolds & Curtin, 2011). They meet the screening criteria, but show low levels of activity and do not seem to be very serious about taking their start-up idea to the market (or to termination) in follow-up interviews.

Variables

In addition to the two dependent variables firm birth and firm abandonment, 40 independent variables were included in our modeling (Appendix A). There are six broad types of independent variables: (1) personal characteristics (for example, gender, age, education, start-up experience, industry experience, management experience, (2) motivational elements (for example, motivation to start the business, growth preference), (3) venture related characteristics (for example, team size, ownership structure, industry, hi-tech venture), (4) start-up activities (for example, business plan prepared, outside funding received, own money invested into the business, patent or trademark applied, employees or managers hired), (5) a market related measure (crises), and (6) four complexity measures (rate, concentration, timing, and effort) which we explain hereafter. We did not include any variables measuring social capital because they were not available in our data set.

Rate is defined as the total number of start-up activities undertaken by the nascent entrepreneur divided by the duration of the gestation process of the new business. For example, if the nascent entrepreneur has conducted five different activities in month 1, 3, 6, 9, and 12. This sequence would equal five activities conducted within a time span of 12 months, resulting in a rate of .42, thus reflecting the average pace of organizing activities across the gestation process.

Concentration reflects how closely start-up activities are undertaken in relation to each other. It is operationalized in terms of the variance of monthly activity time. High values reflect a high dispersion of activities whereas low values indicate that more of the start-up activities are bundled together (for example, variance = 0 if all activities are conducted in one month). For example, cases with a start-up activity sequence of {1, 1, 1, 1} and {1, 3, 6, 6} have a concentration of 0 and 6 respectively.

Timing indicates whether the bulk of start-up activities is accomplished earlier or later during the start-up process. It is measured by taking the average event time divided by the duration of the gestation process. For example, the

average event time related to the start-up activities {1, 3, 3, 6, 12, 12} with a duration of 12 months is 4.5. This figure is divided by the duration of 12 months, resulting in a timing of .375. Values close to 1 indicate that start-up activities occurred at the end of the gestation process whereas values close to 0 mean that activities occurred in the first months of the gestation process.

Effort captures the development of the start-up activities over the 60-month period. It is calculated by computing the difference between two periods for total amount of conducted activities in each period. For example, if the entrepreneur conducted 2 activities at t_0, which is the minimum number of activities to be considered nascent, and 5 activities six months later at t_1, the effort invested over this period is calculated by subtracting ($t_1 - t_0$) the number of activities divided by the duration, that is 6 months. The higher the value, the higher the effort in the time sequence.

The data set consists of numerical (for example, team size, rate, concentration, timing, and effort of the conducted start-up activities) as well as categorical (for example, education, motivation, and so on) variables. Start-up activities (for example, writing a business plan) are coded binary (0 = not conducted; 1 = conducted). External economic conditions such as dotcom and financial crises are considered, with a binary variable (0 = noncrisis year, 1 = crisis year).

Empirical methods

The goal of this study is to predict the likelihood of firm birth and firm abandonment over a 60-month gestation period using different machine learning techniques. The computation is based on independent variables after 12 months (t_1), 24 months (t_2), 36 months (t_3), and 48 months (t_4). The dependent variable firm birth was coded as 1 = firm birth and 0 = otherwise, and firm abandonment was coded 1 = firm abandonment and 0 = otherwise.

We followed three steps to conduct our analysis: (1) data preprocessing as previously outlined, (2) optimization, and (3) evaluation. The optimization phase included selecting the optimal set of independent variables as well as conducting hyperparameter tuning for each applied technique using either a grid search (GridSearchCV) or a random search (RandomSearchCV) approach with k-fold cross-validation (k = 10). Both approaches are common techniques to optimize the models hyperparameters and to derive an optimized model (Bergstra & Bengio, 2012; Vo et al., 2019). The choice of the approach was based on the computational power needed (that is, random search for k-nearest neighbors, decision tree, random forest, XGBoost, support vector machine, artificial neural networks, and grid search for the logistic regression). The evaluation step included the testing process with the model comparison. An overview of the methodology is shown in Figure 1.

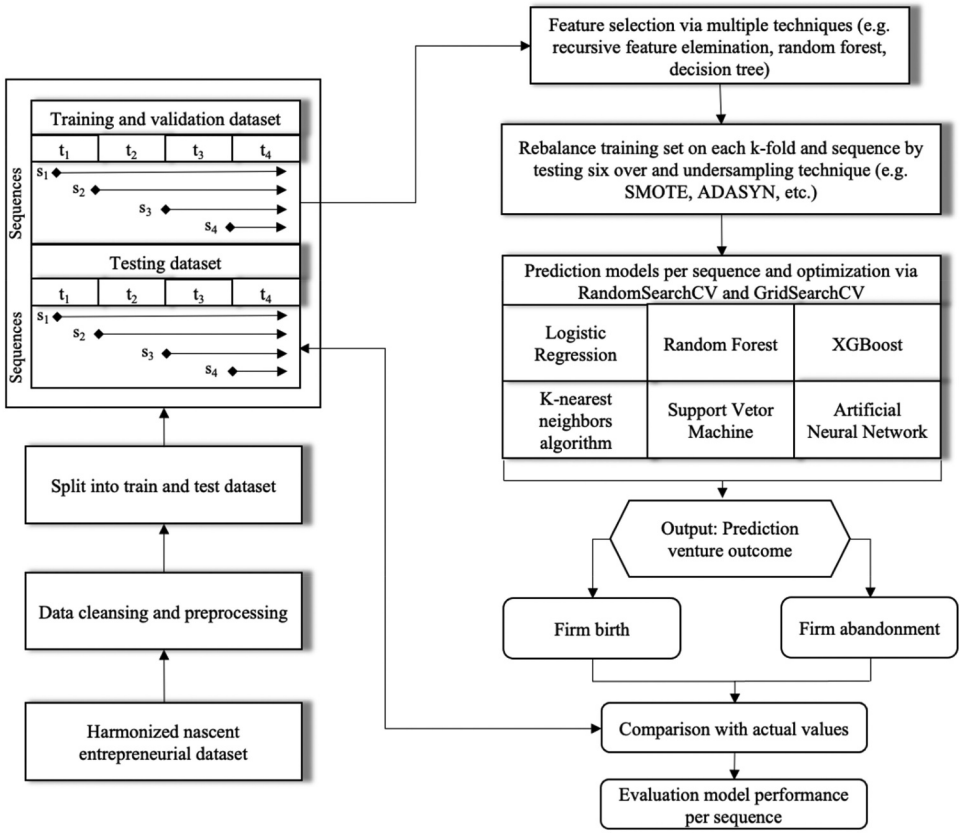

Figure 1. Schematic outline of the analytical approach.

After the data preprocessing, the initial step was to split the data into training, validation, and testing data sets. Training and validation included 75% of the observations (n = 1089) while 25% of the data set are used as test and hold-out set (n = 363) to validate the performance of the models. Given the imbalanced data set, the data split included stratification based on the outcome variable.

To further optimize the models, different variable combinations were tested to identify better performing models. In so doing, we applied seven different feature selection methods: Pearson correlation coefficient, chi-squared, random forest, decision tree, lasso regression, XGBoost and recursive feature elimination (RFE) using three different classifiers (logistic regression, k-nearest neighbors, support vector machine). A feature was considered important if it was selected in seven out of the nine applied techniques. As one of the goals in this study was to analyze the differences in feature importance between the time sequences (t_1–t_4), this procedure was repeated four times for each set of independent variables.

Following the features selection, the train and validation procedure builds on a k-fold cross-validation step (k = 10) to compute the highest F1-score for each model and time sequence. This cross-validation procedure ensures that any sampling biases can be eliminated from the training process (Topuz et al., 2018). To ensure that the selected features improve the model performance, a comparison of the mean values from each k-fold of the full and feature-selected model was conducted.

As we aimed to identify entrepreneurs who are more likely to achieve firm birth or to abandon the venture, the minority classes of firm birth and firm abandonment needed to be accurate. Thus, during the training process of the models, we applied four class balancing techniques (Burnaev et al., 2015) to put more weight on the firm birth and firm abandonment predictions: synthetic minority oversampling technique (SMOTE; Han et al., 2005), adaptive synthetic sampling (ADASYN), neighborhood cleaning rule (Laurikkala, 2002), and edited nearest neighbor (Raniszewski, 2010). Depending on the model, we applied the technique that increased the F1-score and provided reasonable metrics for the area under the receiver operating curve and accuracy (ROC AUC), while keeping the recall or precision metrics on a relative acceptable level. The F1-score is then defined as the harmonic mean of precision and recall. The closer the value to 1, the better the model.

We used three criteria to evaluate the models: (1) accuracy for the probability of correct classifications, (2) precision as well as recall while using a confusion matrix for recognizing the nascent entrepreneur of a certain group (that is, firm birth or firm abandonment), and (3) ROC AUC and F1-score for evaluating the overall performance (Topuz et al., 2018; Veganzones & Séverin, 2018). These measures are calculated for each k-fold (k = 10), averaged for each classification technique and sequence to obtain an overall estimate of the performance of the model. Finally, all trained models were evaluated with the hold-out, test data set.

Classification methods

This paper tackles a typical classification problem using an imbalanced data set. We used several classification methods, including classical logistic regression and five machine learning methods (k-nearest neighbors, random forest, XGBoost, support vector machine, artificial neural network) to predict the likelihood of firm birth and firm abandonment.

K-nearest neighbors' algorithm

Popular in pattern recognition and to solve classification problems, the k-nearest neighbors' (k-NN) algorithm is considered an efficient and relatively simple, easy-to-implement supervised machine learning algorithm (Wang

et al., 2013). The algorithm is based on the concept that data points of the same class should be closer in the feature space. The distance can be defined via the number of samples closest in distance to the new, prediction point (k-nearest neighbor learning). Due to its simplicity and effectiveness, this algorithm has been applied in a variety of settings and has generally provided robust results (Li & Wang, 2015).

Using RandomSearchCV, our configuration included the number of neighbors ranging from 2–70 in steps of two, four different algorithm types (auto, ball tree, k-d tree, and brute force), leaf size in steps of 5 ranging from 10–40, as well as weights (uniform or distance) and a power parameter (1 or 2).

Random forest

This technique consists of a large number of individual decision trees that operate as an ensemble and overcome weaknesses of simple decision trees, such as high sensitivity to small variations in data (Loureiro et al., 2018). Each tree is grown using a random subset of the input variables and at each split a random sample of predictors is examined. The tree is then allowed to grow fully. Thus, no pruning techniques are required. In addition, RF is very user friendly as it requires the researcher to determine two main parameters (that is, the number of variables used for building the individual trees and the number of trees) (Antretter et al. 2019). Random forest has provided good performance in recent studies in entrepreneurship (Sabahi & Parast, 2020; Xu et al., 2018).

The number of trees to grow were tested for the values 10, 50, 100, 150, 200, 250, 500, 1000, 1500, 2000. Values of depth ranged from 2–50 in steps of 2, while the splits ranged from 10–200 in steps of 25.

Extreme gradient boosting

Extreme gradient boosting (XGBoost) is an efficient implementation of the gradient boosting approach (Friedman, 2001). Simply put, while boosting refers to modifying weak learners (that is, decision trees) to strong learners, the objective of gradient boosting is to minimize weakness based on a gradient decent approach, addressing the loss of the model by adding weak learners. XGBoost is an improved model that introduces a regularized model formalization, thus reducing overfitting and increasing the predictive performance. Among others, XGBoost has been applied to predict new venture creation (Antretter et al. 2019) and survival (Climent et al., 2019).

We tuned a variety of parameters to compute XGBoost. A “gbtree” booster was applied and the parameter tuned to include the number of trees (500, 1000), depth (3, 5, 7, 9), learning rate (.01, .1, .2, .3), gamma (0–.4) and subsampling with values ranging from .5 to .9 with steps of .1.

Support vector machine

Support vector machine (SVM) uses a subset of training points in the decision function (called support vectors), so it is also memory efficient (Kraus & Feuerriegel, 2017). Because of the relative simplicity and flexibility for addressing a range of classification problems, SVMs have been effective even with limited sample sizes (Tu et al., 2019) and often outperform other classical statistical models (Chaudhuri & Bose, 2020). This learning method has been widely applied in different research fields, including the entrepreneurship literature (Blanco-Oliver et al., 2014; Tu et al., 2019), to predict a variety of outcomes such as credit rating (Huang et al., 2004) and financial distress (Blanco-Oliver et al., 2014).

Specifically, we used RandomSearchCV to examine the influence of two different kernel types (linear, rbf) as well as the cost and gamma parameters required for each kernel. For the parameter cost, values ranging from .1–2 with a step of were considered, for gamma, next to auto and scale, the boundary test values ranged from .5–5 with steps of .5.

Artificial neural network

We used a multilayer perceptron (MLP) model with two hidden layers drawing on backpropagation learning methods. Backpropagation refers to the training and learning process of a neural network and is currently one of the most widely used neural network algorithms (Huang et al., 2004; LeCun et al., 2015). Note that an artificial neural network (ANN) with more than one layer is often considered a deep neural network. ANNs have been previously applied in business contexts such as bankruptcy prediction models (Veganzones & Séverin, 2018) and offer several useful properties and capabilities such as nonlinearity, learning from examples, adaptivity and fault tolerance (Friedman, 2001).

For the MLP model, the configuration of the parameters was based on a random search k-fold (k = 10) approach. It included the number of hidden neurons ranging between 25 and 200 with a step of 25 for each hidden layer, different activation functions (relu, tanh, sigmoid, hard sigmoid, swish), different solvers (sgd, adam, nadam, adagrad, adadelta, rmsprop and lfbgs), with iterations ranging from 25–300 with steps of 25, batch sizes ranging from 10–150 with steps of 10, and three different learning rates (constant, adaptive, and in-scaling).

Results

Comparing prediction models

Table 2 summarizes the results for each prediction model and sequence (t_1–t_4) of the new venture gestation. When being benchmarked against a simulated random classification, all models significantly outperform that baseline over all

Table 2. Evaluation of classification models on test data set.

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	54.,27%	66.12%	77.69%	72.73%
	ROC AUC	62,.48%	73.60%	79.44%	82.71%
	Precision	28.57%	38.27%	51.67%	45.39%
	Recall	63.53%	72.94%	72.94%	81.18%
	F1 Score	39.42%	50.20%	60.49%	58.23%
	TP/TN/FP/FN**	54/143/135/31	62/178/100/23	62/220/58/23	69/195/83/16
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	RUS	ROS	EDN	EDN
	Hyperparameters	C = 4, penalty = "l2," solver = "lbfgs"	C = 02, penalty = "l1," solver = "saga"	C = .1, penalty = "l1," solver = "saga"	C = .1, penalty = "l1," solver = "liblinear"
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	49.59%	60.61%	61.16%	71.07%
	ROC AUC	62.33%	62.66%	64.13%	72.68%
	Precision	40.01%	46.01%	46.10%	58.39%
	Recall	82.31%	57.69%	50.00%	66.92%
	F1 Score	53.90%	51.19%	47.97%	62.37%
	TP / TN / FP / FN**	107/73/160/23	75/145/88/55	65/157/76/65	87/171/62/43
Optimization	Nr. Features	31	27	25	62
	Sampling Technique*	EDN	ADA	ADA	ADA
	Hyperparameters	C = 2, penalty = "l1," solver = "liblinear"	C = .1, penalty = "l2," solver = "saga"	C = .2, penalty = "l2," solver = "saga"	C = .3, penalty = "l1," solver = "liblinear"
Model Classifier: K-nearest neighbors (KNN)					
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	55.65%	73.00%	73.55%	77.96%
	ROC AUC	69.19%	74.54%	79.28%	82.67%
	Precision	30.41%	44.25%	45.80%	52.10%
	Recall	69.41%	58.82%	70.59%	72.94%
	F1 Score	42.29%	50.51%	55.56%	60.78%
	TP/TN/FP/FN**	59/143/135/26	50/215/63/35	60/207/71/25	62/221/57/23
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	ADA	ROS	RUS	NCR

(*Continued*)

Table 2. (Continued).

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
	Hyperparameters	n_neighbors = 48, weights = "uniform," algorithm = "brute," leaf_size = 15, $p = 2$	n_neighbors = 68, weights = "uniform," algorithm = "brute," leaf_size = 35, $p = 1$	n_neighbors = 44, weights = "uniform," algorithm = "brute," leaf_size = 25, $p = 1$	n_neighbors = 68, weights = "uniform," algorithm = "brute," leaf_size = 25, $p = 1$
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	45.45%	60.06%	58.13%	61.43%
	ROC AUC	57.92%	63.26%	63.25%	68.46%
	Precision	37.86%	46.11%	44.44%	47.47%
	Recall	81.54%	68.46%	67.69%	72.31%
	F1 Score	51.71%	55.11%	53.66%	57.32%
	TP/TN/FP/FN**	106/59/174/24	89/129/104/41	88/123/110/42	94/129/104/36
Optimization	Nr. Features	31	27	25	22
	Sampling Technique*	EDN	RUS	RUS	SMT
	Hyperparameters	n_neighbors = 6, weights = "distance," algorithm = "kd_tree," leaf_size = 25, $p = 2$	n_neighbors = 64, weights = "distance," algorithm = "brute," leaf_size = 25, $p = 1$	n_neighbors = 68, weights = "uniform," algorithm = "brute," leaf_size = 35, $p = 1$	n_neighbors = 68, weights = "uniform," algorithm = "brute," leaf_size = 35, $p = 1$
Model Classifier: Random Forest (RF)					
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	56.20%	46.01%	59.78.%	58.95%
	ROC AUC	59.40%	59.11%	62.84.%	70.24%
	Precision	42.49%	39.29%	45.56%	45.85%
	Recall	63.08%	93.08%	63.08%	80.77%
	F1 Score	50.77%	55.25%	52.90%	58.50%
	TP/TN/FP/FN**	82/122/111/48	121/46/187/9	82/135/98/48	105/109/124/25
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	ADA	EDN	ROS	NCR
	Hyperparameters	n_estimators = 2000, max_depth = 2, min_samples_split = 50, max_features = "sqrt"	n_estimators = 100, max_depth = 18, min_samples_split =150, max_features = "sqrt"	n_estimators = 200, max_depth = 46, min_samples_split = 150, max_features = "auto"	n_estimators = 200, max_depth = 30, min_samples_split = 100, max_features = "sqrt"

(*Continued*)

Table 2. (Continued).

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	44.63%	61.71%	57.85%	61.71%
	ROC AUC	62.93%	63.05%	62.24%	66.34%
	Precision	38.66%	46.94%	43.58%	47.49%
	Recall	93.08%	53.08%	60.00%	65.38%
	F1 Score	54.63%	49.82%	50.49%	55.02%
	TP/TN/FP/FN**	121/41/192/9	69/155/78/61	78/132/101/52	85/139/94/45
Optimization	Nr. Features	31	27	25	22
	Sampling Technique*	EDN	ROS	ROS	RUS
	Hyperparameters	n_estimators = 150, max_depth = 34, min_samples_split = 10, max_features = "auto"	n_estimators = 50, max_depth = 36, min_samples_split = 175, max_features = "auto"	n_estimators = 50, max_depth = 6, min_samples_split = 175, max_features = "auto"	n_estimators = 100, max_depth = 18, min_samples_split = 10, max_features = "sqrt"
Model Classifier: eXtreme Gradient Boosting (XGBoost)					
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	72.18%	74.38%	78.79%	75.48%
	ROC AUC	76.71%	77.91%	83.39%	83.08%
	Precision	43.75%	46.61%	53.33%	48.48%
	Recall	65.88%	64.71%	75.29%	75.29%
	F1 Score	52.58%	54.19%	62.44%	58.99%
	TP/TN/FP/FN**	56/206/72/29	55/215/63/30	64/222/56/21	85/210/68/21
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	EDN	EDN	NCR	EDN
	Hyperparameters	learning_rate = .2, n_estimators = 400, max_depth = 3, min_child_weight = 2, gamma = .5, subsample = .7, colsample_bytree = .2	learning_rate = .3, n_estimators = 100, max_depth = 7, min_child_weight = .5, gamma = .9, subsample = .8, colsample_bytree = .1	learning_rat = .01, n_estimators = 200, max_depth = 7, min_child_weight = 1, gamma = .6, subsample = .5, colsample_bytree = .8	learning_rate = .01, n_estimators = 400, max_depth = 8, min_child_weight = 3, gamma = .8, subsample = .9, colsample_bytree = .2
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$

(*Continued*)

Table 2. (Continued).

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	51.79%	60.33%	52.89%	65.56%
	ROC AUC	61.41%	60.73%	62.98%	68.31%
	Precision	41.31%	44.93%	41.63%	51.82%
	Recall	82.31%	47.69%	78.46%	54.62%
	F1 Score	55.01%	46.27%	54.40%	53.18%
	TP/TN/FP/FN**	107/81/152/23	62/157/76/68	102/90/143/28	71/167/66/59
Optimization	Nr. Features	31	27	25	22
	Sampling Technique*	EDN	ROS	EDN	SMT
	Hyperparameters	learning_rate = .2, n_estimators = 600, max_depth = 8, min_child_weight = .5, gamma = .7, subsample = .5, colsample_bytree = .8	learning_rate = .1, n_estimators = 200, max_depth = 2, min_child_weight = .5, gamma = 0., subsample = 0., colsample_bytree = .9	learning_rate = .1, n_estimators = 200, max_depth = 2, min_child_weight = 4, gamma = 0., subsample = .5, colsample_bytree = .4	learning_rate = .1, n_estimators = 200, max_depth = 2, min_child_weight = .5, gamma = .7, subsample = .6, colsample_bytree = .9
Model Classifier: Support Vector Machine (SVM)					
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	61.43%	74.66%	77.69%	74.38%
	ROC AUC	63.20%	73.97%	80.18%	82.63%
	Precision	30.77%	46.79%	51.64%	47.18%
	Recall	51.76%	60.00%	74.12%	78.82%
	F1 Score	38.60%	52.58%	60.87%	59.03%
	TP/TN/FP/FN**	44/179/99/41	51/220/58/34	63/219/59/22	67/203/75/18
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	ROS	EDN	NCR	EDN
	Hyperparameters	kernel = "linear," shrinking = True, gamma = "auto," probability = True, C = .5, decision_function_shape = "ovo"	kernel = "linear," shrinking = False, gamma = 2, probability = True, C = .2, decision_function_shape = "ovr"	kernel = "linear," shrinking = False, gamma = 3, probability = True, C = 1.6, decision_function_shape = "ovo"	kernel = "linear," shrinking = True, gamma = 3, probability = True, C = 1.4, decision_function_shape = "ovo"
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	47.66%	52.62%	57.30%	63.91%
	ROC AUC	61.40%	62.97%	67.41%	69.14%

(*Continued*)

Table 2. (Continued).

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
	Precision	39.05%	41.18%	45.58%	49.66%
	Recall	82.31%	75.38%	76.30%	55.38%
	F1 Score	52.97%	53.26%	57.06%	52.36%
	TP/TN/FP/FN**	107/66/167/23	98/93/140/32	103/105/123/23	72/160/73/58
Optimization	Nr. Features	31	27	25	22
	Sampling Technique*	EDN	EDN	RUS	SMT
	Hyperparameters	kernel = "linear," shrinking = False, gamma = 3, probability = True, C = 1.1, decision_function_shape = "ovr"	kernel = "linear," shrinking = False, gamma = 3, probability = True, C = 1.6, decision_function_shape = "ovo"	kernel = "linear," shrinking = False, gamma = 2, probability = True, C = 1.7, decision_function_shape = "ovo"	kernel = "linear," shrinking = True, gamma = 3, probability = True, C = 1.4, decision_function_shape = "ovo"
Model Classifier: Artificial Neural Networks (ANN)					
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	70.25%	65.29%	72.45%	71.63%
	ROC AUC	69.59%	76.09%	80.43%	82.23%
	Precision	39.45%	37.87%	45.03%	44.30%
	Recall	50.59%	75.29%	80.00%	82.35%
	F1 Score	44.33%	50.39%	57.63%	57.61%
	TP/TN/FP/FN**	43/212/66/42	64/173/105/21	68/195/83/17	70/190/88/15
Optimization	Nr. Features	27	25	17	14
	Sampling Technique*	ADA	ROS	NCR	EDN
	Hyperparameters	hidden_layer_sizes = (100,75), max_iter = 150, batch_size = 70, activation = "relu," solver = "sgd," alpha = .0001, learning_rate = "constant"	hidden_layer_sizes = (100,75), max_iter = 150, batch_size = 70, activation = "relu," solver = "sgd," alpha = .0001, learning_rate = "constant"	hidden_layer_sizes = (125,100), max_iter = 150, batch_size = 140, activation = "relu," solver = "sgd," alpha = .4, learning_rate = "adaptive"	hidden_layer_sizes = (100,75), max_iter = 150, batch_size = 70, activation = "relu," solver = "sgd," alpha = .0001, learning_rate = "constant"
Firm abandonment	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
Evaluation metric	Accuracy	47.11%	63.64%	57.02%	61.16%
	ROC AUC	56.96%	63.86%	64.40%	69.78%
	Precision	38.69%	49.31%	43.93%	47.51%
	Recall	81.54%	54.62%	72.31%	80.77%

(*Continued*)

Table 2. (Continued).

		Model Classifier: Logistic Regression (LogR)			
Firm birth	Time	$t_{12} - t_{60}$	$t_{24} - t_{60}$	$t_{36} - t_{60}$	$t_{48} - t_{60}$
	F1 Score	52.48%	51.82%	54.65%	59.83%
	TP/TN/FP/FN**	106/65/168/24	115/68/165/15	94/113/120/36	105/117/116/25
Optimization	Nr. Features	31	27	25	22
	Sampling Technique*	EDN	ROS	NCR	NCR
	Hyperparameters	hidden_layer_sizes = (75,75), max_iter = 250, activation = "relu," batch_size = 150, solver = "adam," alpha = .2, learning_rate = "invscaling"	hidden_layer_sizes = (125,100), max_iter = 150,batch_size = 140, activation = "relu," solver = "sgd," alpha = .4, learning_rate = "adaptive"	hidden_layer_sizes = (100, 75), max_iter = 150, batch_size = 70, activation = "relu," solver = "sgd," alpha = .0001, learning_rate = "constant"	hidden_layer_sizes = (100,75), max_iter = 150, batch_size = 70 activation = "relu," solver = "sgd," alpha = .0001, learning_rate = "constant"

*ROS = RandomOverSampler; RUS = RandomUnderSampler; EDN = EditedNearestNeighbours; NCR = NeighbourhoodCleaningRule; SMT = SMOTE; ADA = ADASYN
** TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative

the observed periods (for example, ROC AUC > .5). In general, the power to predict firm birth or firm abandonment increases over time, in line with the additional information processed by the models over gestation stages (t_1–t_4). In addition, when using the logistic regression model as a benchmark, the analysis generated several important insights for firm birth and firm abandonment.

For firm birth, logistic regression achieves the lowest predictive power after 12 months (accuracy of 54.27%, a ROC AUC of 62.48% and an F1-score of 39.42%) compared to all machine learning models. Conversely, XGBoost achieves the highest predictive power (accuracy of 72.18%, a ROC AUC of 76.71% and an F1-score of 52.58%) for the same period. The predictive power of the logistic regression model increases over the next periods. However, it is still lower compared to the best performing AI model, XGBoost. Overall, our results reveal that XGBoost is the best model to predict firm birth over the different stages of the venture gestation.

For firm abandonment, different models perform better for each time sequence. For example, after 12 months, XGBoost achieves an accuracy of 52% with a F1-score of 55%. After 24 months, k-NN provides the best results with an accuracy of 60.06% and a F1-score of 55.11%, while SVM achieves an accuracy of 57.30% and a F1-score of 57.06% for the predictors after 36 months. The logistic regression model achieves a comparably good predictability after 48 months with an accuracy of 71% and an F1-score of 62%. The performance of ANN for predicting firm abandonment was in the mid-range compared to all the other models over the different gestation sequences.

Translating these numbers into a practical context, various aspects need to be considered. In this study, we aimed for balanced precision and recall metrics, thus tuning the models in order to increase the overall F1-score and evaluate the models based on a reasonable balance between F1-score, accuracy and ROC AUC. However, from a practical perspective, this tuning decision could vary because a higher precision score for firm birth and a higher recall score to identify firm abandonment may be preferred.

In the case of firm birth, if the model accidentally predicts that an investment into a profitable venture is bad (false negative), a chance to invest is missed. This would generate opportunity losses and losses for future financial gains from the missed investment into a newly profitable business venture. If the model predicts that the entrepreneur achieves firm birth, but they do not (false positive) and abandons the venture, the costs also directly relate to monetary and nonmonetary investments. Thus, in the case of identifying a successful nascent entrepreneur, the model with the higher precision should be preferred.

In this respect, XGBoost identifies 120 entrepreneurs as profitable with a precision of 53.33% after 36 months. Out of these 120 profitable entrepreneurs, 64 entrepreneurs are correctly identified as profitable while 56

entrepreneurs are wrongly classified as profitable, thus they have either abandoned or are still in the gestation process after 60 months. In the same gestation period, the neural network achieves a precision of 45.03% and identified 151 entrepreneurs as profitable, out of which 68 entrepreneurs are identified correctly and 83 entrepreneurs are wrongly classified as profitable. While more profitable cases are identified in the neural network model, the amount of wrongly classified entrepreneurs is higher. Thus, XGBoost achieves a better precision score. It is important to note that although models can be optimized for precision, this optimization will be at the expense of lower recall values.

In the case of firm abandonment, the model with a higher recall should be preferred if the goal is to identify entrepreneurs who do not abandon the venture. In other words, it is acceptable to have more false positives (for example, nascent entrepreneurs who are not abandoning are considered to have abandoned the venture) than false negatives (for example, nascent entrepreneurs who abandon the venture are not identified as such). For example, after 12 months, XGBoost predicts that 104 entrepreneurs do not abandon the venture after 60 months. Out of those, 81 entrepreneurs are correctly identified as such while 23 are wrongly classified as abandoning the venture. This equals to a recall value of 82.31%. In other words, if an investor invests in these 104 cases, only 17.69% of the resources invested were allocated to entrepreneurs who abandoned the venture after a period of five years are lost. ANN achieves a recall value of 81.54% and identifies 89 entrepreneurs who did not abandon the venture. Out of these 89 cases, 65 are correctly classified and 24 entrepreneurs are wrongly classified. Fewer cases are identified as not having abandoned the venture and thus, the XGBoost achieves a better recall score.

Factors leading to firm emergence and firm abandonment

Figure 2 summarizes the 10 most important factors predicting firm birth. The reflected factor importance in the figures is based on the results of random forest (RF), one of the classification methods used in our prediction models. RF is characterized by high robustness against overfitting and has delivered high prediction accuracy in a variety of studies in the context of entrepreneurship (Antretter et al. 2019; Sabahi & Parast, 2020; Xu et al., 2018).

As shown in Figure 2, the number of start-up activities conducted, the complexity dynamics (rate, effort, concentration, and timing of activities) as well as a few specific activities (having achieved first sales, having asked for supplier credit, having a formed a start-up team and hired initial employees) are crucial for firm birth. Comparing the different sequences of the gestation

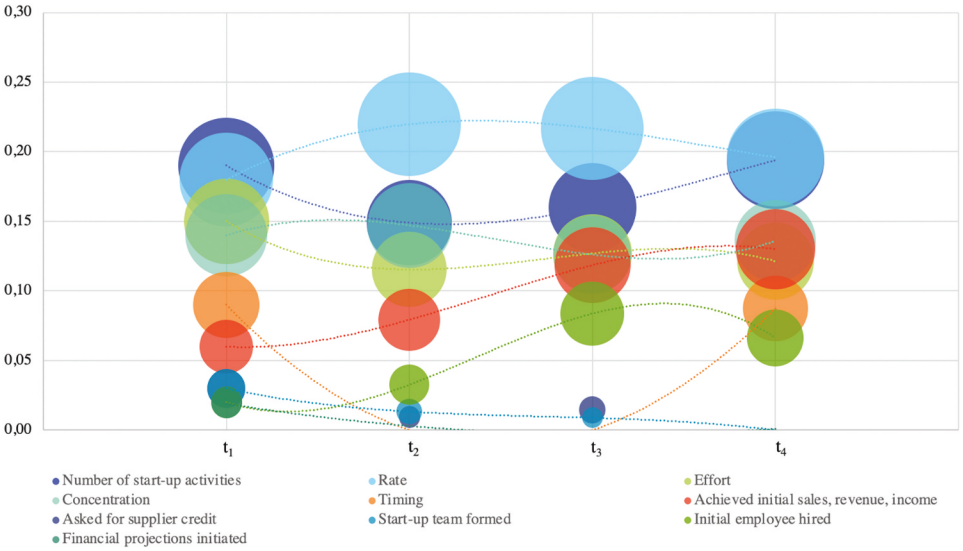

Figure 2. Factor importance of different firm birth prediction sequences.

process, it appears that making financial projections, hiring employees, and having formed a start-up team play a bigger role at an early stage, while asking for supplier credit was important at a later stage (36 and 48 months).

In relation to complexity dynamics, our findings suggest the rate of start-up activities (average pace of organizing), the concentration (extent to which the pace is unstable or constant) as well as the timing (degree to which activities are carried out earlier or later through the process) influence the likelihood of achieving firm birth. Specifically, and in line with previous research (Lichtenstein et al. 2007), we find evidence that a high rate of start-up activities, a minimum pace of activities over time, as well as a tendency to conduct activities later rather than earlier in the process positively relate to firm birth.

We further contribute to the complexity dynamics perspective by adding a new factor, organizing effort, which captures the change in the number of start-up activities over a period. Our results suggest that organizing effort plays a crucial role in predicting firm birth, along with rate, concentration, and timing of activities. Looking at dynamics over time (t_1–t_4), different insights can be generated. First, after 12 months in the gestation process (t_1), all complexity dynamics (rate = 0.18, effort = .15, concentration = .14 and time = .09) play an important role in predicting firm birth. In the following periods (t_2–t_3), the impact of the rate variable (t_3 = .22) to predict firm birth increases while the impact of time (t_3 = .09), concentration (t_3 = .13) and effort (t_3 = .13) remained mostly stable. In the last period, the impact of rate (t_4 = .19) and effort (t_4 = .14) increases again, while the impact of the two other dynamics factors remain stable. From a process perspective, this indicates that complexity

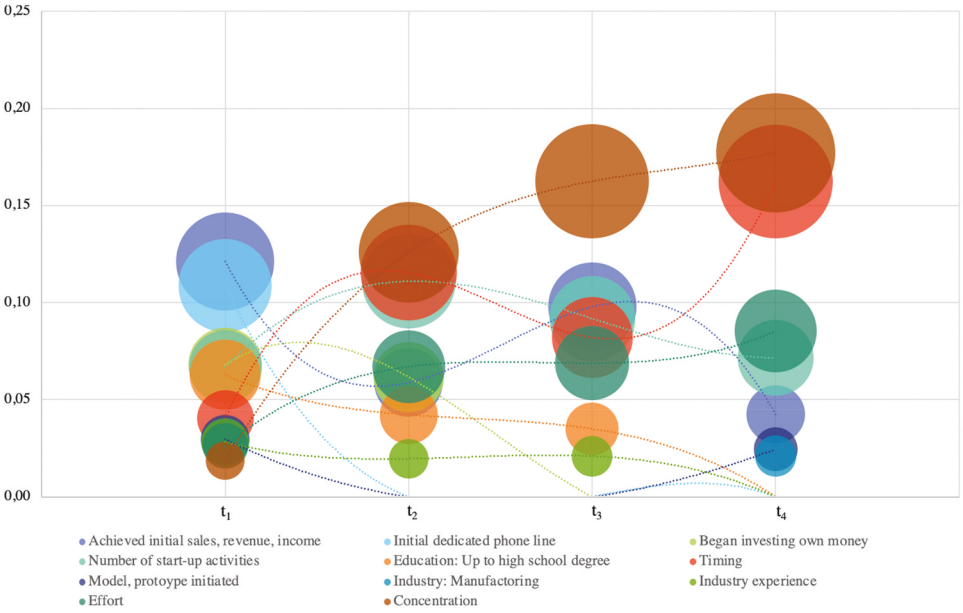

Figure 3. Factor importance of different firm abandonment prediction sequences.

dynamics play a crucial, enduring role throughout the venture creation process. Moreover, the generation of initial sales and team formation variables (that is, initial employees hired) become increasingly important over time.

For firm abandonment (Figure 3), a series of three single organizing activities (achieve initial sales, own money invested, phone lines installed), personal characteristics (educational attainment and industry experience), complexity dynamics, the industry type, and the number of start-up activities conducted appear to be the most important predicting indicators. In terms of timing, key activities such as generating initial sales or having invested the own money as well as certain personal characteristics such as the educational attainment are more important to predict firm abandonment during the earlier stages compared to the complexity dynamics whose importance to predict firm abandonment increases during the later stages. Finally, there are some variables whose importance to predict firm abandonment differ during different sequences.

Discussion

By exploring the combinations of single conditions, and by using machine learning methods to forecast firm emergence, different insights from a methodological, theoretical, and practical perspective can be generated.

From a methodological point of view, our results suggest that machine learning methods significantly outperform a simulated random classification and thus provide a valid option to predict the likelihood of firm birth as well as firm abandonment. In addition, we found evidence that certain machine learning algorithms can, especially during an early stage with ambiguous information, outperform traditional regression-based models in predicting firm birth while preserving interpretability. One explanation for this could relate to the capability of complex models to detect ambiguity of interaction and nonlinear effects in input data (Gerasimovic & Bugaric, 2018), especially when available information for a clear classification at an early stage is scarce. This appears to be especially true for predicting firm birth.

Among the methods we examined, XGBoost was one of the most promising, while neural networks provided comparable performance metrics, suggesting that they can still be used for relatively small data sets. When comparing ANN and XGBoost techniques, we can recognize that XGBoost often achieves state-of-the-art results and outperforms artificial neural networks, especially where data sets are small and structured (Chen & Guestrin, 2016; Climent et al., 2019). However, ANNs include a complex set of hyperparameters that can be tuned and given the applied random search approach, additional optimization to achieve even better results cannot be ruled out. Finally, neural networks often achieve better results if trained on larger data sets and further optimization is conducted using deeper network structures (D'souza et al.'s, 2020). Even though ANN performed worse than XGBoost, ANN is still promising given that, at an early stage of the start-up process (for example, 12 months), most of the entrepreneurs who have not abandoned their business venture are correctly classified.

From a theoretical perspective, our results provide an insight into the critical activities carried out at different stages of the gestation process and a better understanding of the factors leading to firm birth and firm abandonment. For example, our results provide evidence that certain aspects of human capital (that is, education or industry experience) are more prevalent for predicting firm abandonment than firm birth and are more important at an early stage of the gestation process to predict emergence outcomes. While the reasoning for this could be manifold (for example, relationship between human capital and high-velocity decision-making to quit), these insights contribute to existing theories and empirical generalizations related to human capital in entrepreneurship. More pointedly, while some research on the impact of human capital on entrepreneurial outcomes has been inconclusive (Bosma et al., 2004), scholars started to focus on two ways to reflect upon and reconcile this mixed evidence (Dimov, 2017). The first relates to the complexity of different entrepreneurial outcomes which calls for further research on identifying new possible moderators. The second relates to the nature of

the human capital construct itself and the way it is constructed and measured (Dimov, 2017). With our analysis, we add a possible third explanation to this discussion by connecting temporal elements to the relationship between certain indicators of human capital and a specific entrepreneurial outcome (that is, firm abandonment).

In a similar vein, we provide new insights about new venture team formation and the likelihood of achieving firm birth (Held et al., 2018; Klotz et al., 2014). Our results suggest that presence of a founding team is a stronger predictor for firm birth in the first and last stages of the gestation process. It is essential to outline that the terminology "stages" is more related to a temporal dimension used in the context of forecasting firm birth or firm abandonment. When considering the "three stage model" outlined by Davidsson and Gruenhagen (2020, p. 17) for example, the human capital dimensions included in the models can occur in any of the three proposed stages "prospecting, developing or exploiting" and potentially lead firm birth. However, from a temporal perspective, team formation (that is, hiring employees) seems to be more important at either a very early or later stage and probably can indicate, similar to achieving a first sale, a "critical incident" (Davidsson & Gruenhagen, 2020, p. 18) for predicting firm birth.

In addition, our results contribute to the discussion whether entrepreneurship education should be considered as a method or as a process (Neck & Greene, 2011). While the process perspective follows "one of identifying an opportunity, developing the concept, understanding resource requirements, acquiring resources, implementation, and exit" (Neck & Greene, 2011, p. 59), thus having at its core opportunity evaluation, feasibility analysis, business planning, and financial forecasting, the method perspective "represents a body of skills or techniques." (Neck & Greene, 2011, p. 61) Given our research design, at first sight, the nature of our work rather relates to a process perspective with a "planning and prediction" character using AI. One can reasonably ask: "If AI allows one to predict the likelihood of, for example, firm birth during different stages, should it not inevitably follow a process?" The answer is equivocal. Our results highlight the possibility of critical incidents that allow for a prediction of firm birth either early or later during the process. As such, teaching from a process perspective may align toward getting the best out of such temporally aligned incidents (for example, sales classes). Despite these insights, it is important to note that certain critical incidents can also occur at any time during the process, thus making the mentioned "three stage model" from a process perspective intriguing. Moreover, the most crucial prediction variables relate to complexity dynamics such as rate, time, concentration, or effort. These variables cannot be assigned to any of the stages illustrated in the process paragraph above and are rather to be understood as overshadowing the whole phase. As such, a method approach for teaching these skills can be

beneficial. For example, in relation to the rate, education programs should focus on would-be entrepreneurs should maintain or even increase the pace of entrepreneurial activities, given its importance for firm birth. A design thinking (Linton & Klinton, 2019) or even a design sprint approach (Hilliard, 2021) could help in developing such skills.

From a practical point of view, the combination of all proposed firm birth and firm abandonment prediction models provides a valuable system to foster a better allocation of resources to successful entrepreneurs, while reducing respective resource misallocation to entrepreneurs who abandon the venture. The allocation of third-party resources to potentially successful entrepreneurs is inherently speculative given the high failure rate during the start-up process, the often patchy product or service offerings, and the unproven technologies (Drover et al., 2017). While information is often scarce at such an early stage, uncertainty and information asymmetry prevail (Dunkelberg et al., 2013; Nguyen et al., 2020). This asymmetry can lead to agency problems (Jensen & Meckling, 1976) that arise due to hidden information and hidden actions between the involved parties.

The proposed models provide stakeholders with a viable early stage screening and monitoring system to mitigate agency problems during the venture creation process. Specifically, the models can contribute to not only mitigate costs of a resource misallocation, but also to the costs related to the selection process for external parties such as incubators, accelerators, angel investors and venture capitalists (Drover et al., 2017; Yin & Luo, 2018). For example, a seed venture capitalist must often use more visible and quickly accessible information to efficiently discern which ventures are worthy of moving to due diligence (Drover et al., 2017). Given the considerable amount of resources that are expended in properly vetting the venture during the due diligence (Drover et al., 2017), our models provide a parsimonious solution to distinguish between entrepreneurs who are likely to achieve firm birth and those who are likely to abandon during the screening process.

Moreover, the identification of entrepreneurs who are likely to abandon their project could also help business advisers to timely engage with the entrepreneurs before their abandonment decision may even arises. This could help at an early stage to identify crucial factors to boost the nascent entrepreneurial motivation or to guide them to pivot their idea.

Conclusion

Predicting new venture gestion outcome is a complex endeavor. More accurate forecasting tools can serve as a support and an early warning system to help stakeholders identify promising entrepreneurs, while at the same time provide valuable insights into the gestation process. Against this

backdrop, we used machine learning methods to forecast the likelihood of firm birth and firm abandonment during the first five years of a new business gestation, and to identify what single factors can lead to firm emergence. Our results suggest that the application of AI techniques to predict firm birth and firm abandonment is very promising. We provide evidence that machine-learning algorithms outperform traditional regression-based models while preserving interpretability. Among the methods we examined, XGBoost was one of the most promising, while neural networks provided comparable performance metrics, showing that they can be used for relatively small data sets.

In addition, by identifying key factors to predict firm birth and firm abandonment, we were able to gain valuable insights in relation to the start-up activities leading to firm emergence. Achieving sales, as well as the rate, timing, and concentration of activities are key elements in predicting the venture gestation outcome. Looking at the whole firm gestation process, a variety of activities are more significant at an early stage (for example, forming a start-up team), while others are more important at a later stage of the gestation process (for example, ask for supplier credits) in relation to predicting the outcome of the venture. This dynamism underpins the complexity of the gestation process and further underlines the importance for external parties to gain this in-depth knowledge of the entrepreneur, their status, and their business environment. Combining these elements, we thus contribute to theory and practice by providing further insights, as outlined in the discussion section, to a better understanding of the entrepreneurial process and the practical need for such better understanding (that is, cost related to resource misallocation).

This study has a several limitations. First, it is important to recognize that machine learning methods do not constitute a panacea in decision-making as they are constrained when processing and interpreting "soft" types of information (information that cannot be quantified) and making predictions in uncertain situations (Dellermann et al., 2017). Although these models can provide some guidance, entrepreneurship often requires intuitive decision-making and heuristics enable entrepreneurs to function effectively in those situations. Similarly, experts such as business angels or venture capitalists, with their industry knowledge and sensitivity to entrepreneurial personality, can provide informed advice. Second, firm emergence is a complex process where further variables may play an essential role on the venture's outcome. For nascent entrepreneurs, cognitive capacities can have a significant effect on the likelihood of succeeding (SBA, 2012) and venture networking (that is, connection to incubators, accelerators, research centers, universities; founder/s' strong and weak ties, and so on), may positively impact the venture outcome (Woolley & MacGregor, 2021).

Given the missing data in the used harmonized data set, future research could include such variables into the proposed models. Third, as outlined in the sequential analysis, stakeholders using classification models during the start-up process must be aware of the changes over time. Henceforth, the models need to be retrained at least on a yearly basis to provide the necessary information.

Finally, users should be aware of the complexities associated with the different models, especially regarding their implementation. Given that data-mining techniques such as neural networks and support vector machines are often considered a black box themselves (Cortez & Embrechts, 2013), models with better visualization possibilities such as random forest models, may provide a cost-effective alternative. To improve the predictability of such models, future research would profit from including further variables such as cognitive factors, competitor data, and information about the quality of the business idea and perceived product–market fit. Further model optimizations such as including additional layers in the neural network model should be considered. Combining text mining models, for example, business plans and pitches with machine learning models could further increase the predictive power.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Paris Koumbarakis http://orcid.org/0000-0003-0078-1962

References

Antretter, T., Blohm, I., Grichnik, D., & Wincent, J. (2019). Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights*, *11*, e00109. https://doi.org/10.1016/j.jbvi.2018.e00109

Arenius, P., Engel, Y., & Klyver, K. (2017). No particular action needed? A necessary condition analysis of gestation activities and firm emergence. *Journal of Bussiness Venturing Insights*, *8*, 87–92. https://doi.org/10.1016/j.jbvi.2017.07.004

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Blanco-Oliver, A., Pino-Mejías, R., & Lara-Rubio, J. (2014). Modeling the financial distress of microenterprise start- ups using support vector machines: A case study. *Innovar*, *24*(1Spe), 153–168. https://doi.org/10.15446/innovar.v24n1spe.47615

Bosma, N., Van Praag, M., Thurik, R., & De Wit, G. (2004). The value of human and social capital investments for the business performance of startups. *Small Business Economics*, *23* (3), 227–236. https://doi.org/10.1023/B:SBEJ.0000032032.21192.72

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of resampling on accuracy of imbalanced classification. In A. Verikas, P. Radeva, & D. Nikolaev, (Eds.), *Eighth International Conference on machine vision* (vol. 987521). https://doi.org/10.1117/12.2228523

Chandler, G. N., Honig, B., & Wiklund, J. (2005). Antecedents,mmoderators, and performance consequences of membership change in new venture teams. *Journal of Bussiness Venturing*, *20*(5), 705–725. https://doi.org/10.1016/j.jbusvent.2004.09.001

Chaudhuri, N., & Bose, I. (2020). Exploring the role of deep neural networks for post-disaster decision support. *Decision Support Systems*, *130*, 113234. https://doi.org/10.1016/j.dss.2019.113234

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD International Conference of knowledge discovery and data mining*, ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785.

Chwolka, A., & Raith, M. G. (2012). The value of business planning before start-up - a decision-theoretical perspective. *Journal of Bussiness Venturing*, *27*(3), 385–399. https://doi.org/10.1016/j.jbusvent.2011.01.002

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the eurozone: an extreme gradient boosting approach. *Journal of Bussiness Research*, *101*, 885–896. https://doi.org/10.1016/j.jbusres.2018.11.015

Clough, D. R., Fang, T. P., Vissa, B., & Wu, A. (2019). Turning lead into gold: How do entrepreneurs mobilize resources to exploit opportunities? *Academy of Management Annals*, *13*(1), 240–271. https://doi.org/10.5465/annals.2016.0132

Coad, A., & Srhoj, S. (2019). Catching gazelles with a lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics*, *55*(3), 1–25. https://doi.org/10.1007/s11187-019-00203-3

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Science*, *225*, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

D'souza, R. N., Huang, P. Y., & Yeh, F. C. (2020). Structural analysis and optimization of convolutional neural networks with a small sample size. *Scientific Reports*, *10*(1), 834. https://doi.org/10.1038/s41598-020-57866-2

Davidsson, P., & Honig, B. (2003). The role of social and human capital among nascent entrepreneurs. *Journal of Business Venturing*, *18*(3), 301–331. https://doi.org/10.1016/S0883-9026(02)00097-6

Davidsson, P., & Gordon, S. R. (2012). Panel studies of new venture creation: A methods-focused review and suggestions for future research. *Small Business Economics*, *39*(4), 853–876. https://doi.org/10.1007/s11187-011-9325-8

Davidsson, P., & Gruenhagen, J. H. (2020). Fulfilling the process promise: A review and agenda for new venture creation process research. *Entrepreneurship Theory and Practice*, *45*(5), 1083–1118. https://doi.org/10.1177/1042258720930991

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2017). Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method. *International Conference on Information Systems (ICIS)*, Seoul, South Korea. https://doi.org/10.48550/arXiv.2105.03360

Delmar, F., & Shane, S. (2003). Does business planning facilitate the development of new ventures? *Strategic Management Journal*, *24*(12), 1165–1185. https://doi.org/10.1002/smj.349

Delmar, F., & Shane, S. (2006). Does experience matter? The effect of founding team experience on the survival and sales of newly founded ventures. *Strategic Organization*, *4*(3), 215–247. https://doi.org/10.1177/1476127006066596

Dimov, D. (2010). Nascent entrepreneurs and venture emergence: Opportunity confidence, human capital, and early planning. *Journal of Management Studies*, *47*(6), 1123–1153. https://doi.org/10.1111/j.1467–6486.2009.00874.x.

Dimov, D. (2017). Towards a qualitative understanding of human capital in entrepreneurship research. *International Journal of Entrepreneurial Behaviour and Research*, *23*(2), 210–227. https://doi.org/10.1108/IJEBR-01-2016-0016

Drover, W., Wood, M. S., & Zacharakis, A. (2017). Attributes of angel and crowdfunded investments as determinants of VC screening decisions. *Entrepreneurship: Theory and Practice*, *41*(3), 323–347. https://doi.org/10.1111/etap.12207

Dunkelberg, W., Moore, C., Scott, J., & Stull, W. (2013). Do entrepreneurial goals matter? Resource allocation in new owner-managed firms. *Journal of Business Venturing*, *28*(2), 225–240. https://doi.org/10.1016/j.jbusvent.2012.07.004

Friedman, J. (2001). Greedy function approximation : A gradient boosting machine. *Annals on Stategy*, *29*(5), 1189–1232. https://www.jstor.org/stable/2699986

Gerasimovic, M., & Bugaric, U. (2018). Enrollment management model: Artificial neural networks versus logistic regression. *Applied Artificial Intelligence*, *32*(2), 153–164. https://doi.org/10.1080/08839514.2018.1448146

Giotopoulos, I., Kontolaimou, A., & Tsakanikas, A. (2017). Drivers of high-quality entrepreneurship: What changes did the crisis bring about? *Small Bussiness Economics*, *48*(4), 913–930. https://doi.org/10.1007/s11187-016-9814-x

Hallen, B. L. (2008). The causes and consequences of the initial network positions of new organizations: From whom do entrepreneurs receive investments? *Administrative Science Quarterly*, *53*(4), 685–718. https://doi.org/10.2189/asqu.53.4.685

Hallen, B. L., Cohen, S. L., & Bingham, C. B. (2020). Do accelerators work? If so, how? *Organization Science*, *31*(2), 378–414. https://doi.org/10.1287/orsc.2019.1304

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), *Advances in intelligent compuing* (pp. 878–887). Springer-Verlag.

Hechavarria, D. M., Reno, M., & Matthews, C. H. (2012). The nascent entrepreneurship hub: Goals, entrepreneurial self-efficacy and start-up outcomes. *Small Business Economics*, *39*(3), 685–701. https://doi.org/10.1007/s11187-011-9355-2

Held, L., Herrmann, A. M., & van Mossel, A. (2018). Team formation processes in new ventures. *Small Business Economics*, *51*(2), 441–464. https://doi.org/10.1007/s11187-018-0010-z

Hilliard, R. (2021). Start-up sprint: Providing a small group learning experience in a large group setting. *Journal of Management Education*, *45*(3), 387–403. https://doi.org/10.1177/1052562920948924

Honig, B., & Samuelsson, M. (2012). Planning and the entrepreneur: A longitudinal examination of nascent entrepreneurs in Sweden. *Journal of Small Business Management*, *50*(3), 365–388. https://doi.org/10.1111/j.1540-627X.2012.00357.x

Hopp, C., & Sonderegger, R. (2015). Understanding the dynamics of nascent entrepreneurship-prestart-up experience, intentions, and entrepreneurial success. *Journal of Small Bussiness Management*, *53*(4), 1076–1096. https://doi.org/10.1111/jsbm.12107

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558. https://doi.org/10.1016/S0167-9236(03)00086-1

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Khan, S. A., Tang, J., & Joshi, K. (2014). Disengagement of nascent entrepreneurs from the start-up process. *Journal of Small Business Management*, *52*(1), 39–58. https://doi.org/10.1111/jsbm.12032

Klotz, A. C., Hmieleski, K. M., Bradley, B. H., & Busenitz, L. W. (2014). New venture teams: A review of the literature and roadmap for future research. *Journal of Management*, *40*(1), 226–255. https://doi.org/10.1177/0149206313493325

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38–48. https://doi.org/10.1016/j.dss.2017.10.001

Laurikkala, J. (2002). Instance-based data reduction for improved identification of difficult small classes. *Intelligent Data Anaalysis*, *6*(4), 311–322. https://doi.org/10.3233/ida-2002-6402

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lévesque, M., Obschonka, M., & Nambisan, S. (2020). Pursuing impactful entrepreneurship research using artificial intelligence. *Entrepreneurship Theory and Practice*, 104225872092736. https://doi.org/10.1177/1042258720927369

Li, J., & Wang, Y. (2015). A new fast reduction technique based on binary nearest neighbor tree. *Neurocomputing*, *149*, 1647–1657. https://doi.org/10.1016/j.neucom.2014.08.028

Liao, J. J., & Gartner, W. B. (2007). The influence of pre-venture planning on new venture creation. *Journal of Small Business Strategy*, *18*(2), 1–22.

Liao, J., & Welsch, H. (2008). Patterns of venture gestation process: Exploring the differences between tech and non-tech nascent entrepreneurs. *The Journal of High Technology Management Research*, *19*(2), 103–113. https://doi.org/10.1016/j.hitech.2008.10.003

Liao, J., Welsch, & Moutray, C. (2009). Start-up resources and entrepreneurial discontinuance: The case of nascent entrepreneurs. *Journal of Small Bussiness Strategy*, *19*(2), 89–103. https://libjournals.mtsu.edu/index.php/jsbs/article/view/112

Lichtenstein, B. B., Carter, N. M., Dooley, K. J., & Gartner, W. B. (2007). Complexity dynamics of nascent entrepreneurship. *Journal of Business Venturing*, *22*(2), 236–261. https://doi.org/10.1016/j.jbusvent.2006.06.001

Linton, G., & Klinton, M. (2019). University entrepreneurship education: A design thinking approach to learning. *Journal of Innovation and Entrepreneurship*, *8*(1), 1–11. https://doi.org/10.1186/s13731-018-0098-z

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81–93. https://doi.org/10.1016/j.dss.2018.08.010

Lussier, R. N. (1995). A nonfinancial business success versus failure prediction model for young firms. *Journal of Small Bussiness Management*, *33*(3), 8. https://doi.org/10.1111/j.1540-627X.2010.00298.x

Lussier, R. N., & Claudia, E. H. (2010). A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, *48*(3), 360–377. https://doi.org/10.1111/j.1540-627X.2010.00298.x

Mayr, S., Mitter, C., Kücher, A., & Duller, C. (2020). Entrepreneur characteristics and differences in reasons for business failure: Evidence from bankrupt Austrian SMEs. *Journal of Small Business and Entrepreneurship*, *33*(5), 539–558. https://doi.org/10.1080/08276331.2020.1786647

Muñoz-Bullon, F., Sanchez-Bueno, M. J., & Vos-Saz, A. (2015). Startup team contributions and new firm creation: The role of founding team experience. *Entrepreneurship and Regional Development*, *27*(1–2), 80–105. https://doi.org/10.1080/08985626.2014.999719

Neck, H. M., & Greene, P. G. (2011). Entrepreneurship education: Known worlds and new frontiers. *Journal of Small Business Management*, *49*(1), 55–70. https://doi.org/10.1111/j.1540-627X.2010.00314.x

Newbert, S. L. (2005). New firm formation: A dynamic capability perspective. *Journal of Small Bussiness Management*, *43*(1), 55–77. https://doi.org/10.1111/j.1540-627X.2004.00125.x

Newbert, S. L., & Tornikoski, E. T. (2012). Supporter networks and network growth: A contingency model of organizational emergence. *Small Bussiness Economy*, *39*(1), 141–159. https://doi.org/10.1007/s11187-010-9300-9

Newbert, S. L., Tornikoski, E. T., & Quigley, N. R. (2013). Exploring the evolution of supporter networks in the creation of new organizations. *Journal of Bussiness Venturing*, *28*(2), 281–298. https://doi.org/10.1016/j.jbusvent.2012.09.003

Nguyen, B., Le, C., & Vo, X. V. (2020). The paradox of investment timing in small business: Why do firms invest when it is too late? *Journal of Small Business Management*, 1–43. https://doi.org/10.1080/00472778.2020.1816436

Obschonka, M., & Audretsch, D. B. (2019). Artificial intelligence and big data in entrepreneurship: A new era has begun. *Small Business Economics*, *55*, 529–539. https://doi.org/10.1007/s11187-019-00202-4

Raniszewski, M. (2010). The edited nearest neighbor rule based on the reduced reference set and the consistency criterion. *Biocybernetics and Biomedical Engineering*, *30*, 31–40.

Reynolds, P. D., & Curtin, R. T. (2011). Overview and Commentary. In P. Reynolds & R. Curtin (Eds.), *New business creation: International studies in entrepreneurship* 1 (Vol. 27, pp. 295–334). Springer. https://doi.org/10.1007/978-1-4419-7536-2_11

Reynolds, P. D., Hechavarria, D., Tian, L. R., Samuelsson, M., & Davidsson, P. (2016). *Panel study of entrepreneurial dynamics: A five cohort outcomes harmonized dataset*. http://www.psed.isr.umich.edu/psed/data

Reynolds, P. D. (2017). When is a firm born? Alternative criteria and consequences. *Business Economics*, *52*(1), 41–56. https://doi.org/10.1057/s11369-017-0022-8

Rotefoss, B., & Kolvereid, L. (2005). Aspiring, nascent and fledging entrepreneurs: An investigation of the business start-up process. *Entreprenureship. Reg. Dev*, *17*(2), 109–127. https://doi.org/10.1080/08985620500074049

Sabahi, S., & Parast, M. M. (2020). The impact of entrepreneurship orientation on project performance: A machine learning approach. *International Journal of Production Economics*, *226*, 107621. https://doi.org/10.1016/j.ijpe.2020.107621

SBA. (2012). *Frequently asked questions about small business*. (accessed February 6, 2020) https://www.sba.gov/sites/default/files/FAQ_Sept_2

Shim, J., & Davidsson, P. (2018). Shorter than we thought: The duration of venture creation processes. *Journal of Bussiness Venturing Insights*, *9*, 10–16. https://doi.org/10.1016/j.jbvi.2017.12.003

Steffens, P., Terjesen, S., & Davidsson, P. (2012). Birds of a feather get lost together: New venture team composition and performance. *Small Business Economics*, *39*(3), 727–743. https://doi.org/10.1007/s11187-011-9358-z

Thiess, D., Sir, C., & Grichni, D. (2016). How does heterogeneity in experience influence the performance of nascent venture teams? Insights from the US PSED II study. *Journal of Business Venturing Insights*, *5*, 55–62. https://doi.org/10.1016/j.jbvi.2016.04.001

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A bayesian decision support model. *Decision Support Systems*, *106*, 97–109. https://doi.org/10.1016/j.dss.2017.12.004

Tornikoski, E. T., & Newbert, S. L. (2007). Exploring the determinants of organizational emergence: A legitimacy perspective. *Journal of Business Venturing*, *22*(2), 11–335. https://doi.org/10.1016/j.jbusvent.2005.12.003

Tornikoski, E. T. (2008). Legitimating characteristics and firm emergence. *Journal of Enterprising Culture*, *16*(3), 233–256. https://doi.org/10.1142/S0218495808000144

Tu, J., Lin, A., Chen, H., Lin, Y., & Li, C. (2019). Predict the entrepreneurial intention of fresh graduate students based on an adaptive support vector machine framework. *Mathematical Probems in Engineering*, 1–16. https://doi.org/10.1155/2019/2039872

van Gelderen, M., Thurik, R., & Bosma, N. (2005). Success and risk factors in the pre-startup phase. *Small Bussiness Economics*, *26*(4), 319–335. https://doi.org/10.1007/s11187-004-6837-5

van Witteloostuijn, A., & Kolkman, D. (2019). Is firm growth random? A machine learning perspective. *Journal of Business Venturing Insights*, *11*, 1–5. https://doi.org/10.1016/j.jbvi.2018.e00107

Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, *112*, 111–124. https://doi.org/10.1016/j.dss.2018.06.011

Vegetti, F., & Adăscăliței, D. (2017). the impact of the economic crisis on latent and early entrepreneurship in Europe. *International Entrepreunership Management Journal*, *13*(4), 1289–1314. https://doi.org/10.1007/s11365-017-0456-5

Vo, N. N. Y., He, X., Liu, S., & Xu, G. (2019). Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decision Support Systems*, *124*, 113097. https://doi.org/10.1016/j.dss.2019.113097

Wang, J. S., Lin, C. W., & Yang, Y. T. C. (2013). A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition. *Neurocomputing*, *116*, 136–143. https://doi.org/10.1016/j.neucom.2011.10.047

WBAF. (2020). *Global fundraising stage - GFRS 2020 an international co-investment platform*. World Business Angels Investment Forum. https://www.wbaforum.org/upload/07GFRS_2020_745.pdf

Weinblat, J. (2018). Forecasting European high-growth Firms - A random forest approach. *Journal of Industry, Competition and Trade*, *18*(3), 253–294. https://doi.org/10.1007/s10842-017-0257-0

Wennberg, K., & Anderson, B. S. (2020). Editorial: Enhancing the exploration and communication of quantitative entrepreneurship research. *Journal of Business Venturing*, *35*(3), 1–11. https://doi.org/10.1016/j.jbusvent.2019.05.002

Woolley, J. L., & MacGregor, N. (2021). The influence of incubator and accelerator participation on nanotechnology venture success. *Entrepreneurship: Theory and Practice*. https://doi.org/10.1177/10422587211024510

Xu, B., Yang, J., & Sun, B. (2018). A nonparametric decision approach for entrepreneurship. *International Entrepreneurship and Management Journal*, *14*(1), 5–14.

Yin, B., & Luo, J. (2018). How do accelerators select startups? Shifting decision criteria across stages. *IEEE Transactions on Engineering Management*, *65*(4), 574–589. https://doi.org/10.1109/TEM.2018.2791501

Appendix A.

List of variables included in the study.

Variable	Type of variable	Type of variable
Gender	PC	Categorical
Age	PC	Categorical
Education	PC	Categorical
Born in the country	PC	Categorical
Start-up experience	PC	Categorical
Industry experience	PC	Categorical
Management experience	PC	Categorical
Motivation to start a business	ME	Categorical
Growth preference	ME	Categorical
Team size	VC	Numerical
Ownership (minority, equal, majority)	VC	Categorical
Industry type (manufacturing, professional services, IT, construction, transportation, and agriculture)	VC	Categorical
Hi-tech venture	VC	Categorical
Availability of technology five years ago	VC	Categorical
R&D focus of the venture	VC	Categorical
A phone listing for the business acquired	SA	Categorical
Outside funding from institutions or people received	SA	Categorical
Asked financial institutions or other people for funding	SA	Categorical
Credit from suppliers established	SA	Categorical
Own money invested into the business	SA	Categorical
Patent, trademark, or copyright applied	SA	Categorical
Employees or managers hired	SA	Categorical
Major items equipment or property purchased or rented	SA	Categorical
Material, supplies, inventory purchased	SA	Categorical
New firm registered	SA	Categorical
Devoted full time to business	SA	Categorical
Marketing or promotion activities started	SA	Categorical
Effort made to define the market opportunity	SA	Categorical
Start-up team organized	SA	Categorical
Financial projections prepared	SA	Categorical
Business plan prepared	SA	Categorical
Worked on a model or prototype for product delivery	SA	Categorical
Achieved initial sales	SA	Categorical
Sum of conducted activities for each period (t_1–t_4)	SA	Numerical
Rate	CM	Numerical
Concentration	CM	Numerical
Timing	CM	Numerical
Effort	CM	Numerical
Crises	MA	Categorical

Note: PC = personal characteristics, ME = motivational elements, VC = venture related characteristics, SA = start-up activities, CM = complexity measure, MA = market-related measure